compositionalMS

Jonathon O'Brien

2018-10-22

The compositionalMS package performs analyses on TMT/iTRAQ data under the assumption that all of the relevant quantitative information is captured in the signal proportions (O'Brien et al. 2018). The study of constrained data is the primary focus of compositional data analysis (Aitchison 1986) and the models used in this software are based on prinicples from this field.

A D dimensional composition is defined as any vector $\mathbf{x}$ of non-negative elements $x_1, ..., x_D$ such that $x_1 + ... + x_D = K$.

The closure of a composition is the function that converts a composition into a set of proportions

$$cl(\mathbf{x}) = \mathbf{\frac{x_1}{K}, ..., \frac{x_D}{K}}$$

Normalization

Row Normalization

The models used in this software anticipate singal-to-noise (SN) ratios, which are theoretically greater than or equal to one. This is important because the mathematical foundation of compositional data analysis does not allow components to take a value of zero. While SN ratios should not be less than one, various software packages will occasionally estimate signals below the noise. Accordingly, the first step in our normalization is to replace all SN values which are less than one with a 1.

Each scan is then reduced to its proportional information through the additive log-ratio (ALR) transformation which maps a D dimensional simplex into a vector in $\mathbb{R}^{D-1}$.

$$alr(\mathbf{x}) = \mathbf{y} = \mathbf{\log_2(\frac{x_2}{x_1}), ..., \log_2(\frac{x_D}{x_1})}$$

Notice that the ALR is typically defined with the natural logarithm, but we have defined it here as a log base 2 transformation to match the conventional log-ratio scale used throughout the field of proteomics.

A transformation back into a the simplex (contrained to one), can be achieved with the ALR inverse

$$alr^{-1}(\mathbf{y}) = \mathbf{cl(2^0, 2^{y_1}, ...2^{y_{D-1}})}$$

The first component in the ALR transformation plays a special role as a common reference for all of the log-ratios. Our program treats whatever condition is provided in the first column as the reference. In experiments that include a bridge channel it is natural to use the bridge as the reference. When a bridge is not present, or in any case where the first condition has more than one replicate, all columns belonging to that first condition are collapsed into one column by taking the geometric mean. Notice that converting each scan into additive log-ratios with respect to a bridge channel provides a solution to the cross-plex normalization problem.

Column Normalization

In a mass spectrometry proteomics experiment it is common to assume that the average protein abundance across samples should be equivalent. Typically normalization is done by finding multiplicative factors such that the mean (or sum) of each column is forced to be equivalent. This should compensate for pipetting errors, variation in precipitation and other factors that systematically shift the abundance of all proteins in a sample.

If the function parameters include 'normalize = TRUE' then a similar normalization occurs that is slightly modified to use operations defined in compositional data analysis. Specifically, each column is multiplied so that the geometric means of the columns will be equivalent.

Column and row normalized matrices containing proportions are stored for each plex. These are the values that are plotted when clicking on a specific biological replicate. They should be interpreted as the proportion of signal in the relevant plex that belonged to the peptide shown.

Model Specification

Stastical modeling is performed on the normalized, ALR transformed data. The model is defined as follows:

$$y_{ijk} \sim N(\beta_{ij}, \sigma_{ij})\beta_{ij} \sim N(0, 10)\sigma_{ij} \sim InverseGamma(2, \tau)\tau \sim halfNormal(0, 5)$$

where y_{ijk} is the additive log-ratio, from the reference condition, of the k'th petpide ($k = 1, ..., n_{ij}$) nested within the i'th protein ($i = 1, ..., N$) in the j'th condition ($j = 1, ..., M$). β_{ij} represents the log-ratio (fold-change) of the i'th protein from the reference to the j'th condition. σ_{ij} is the prior standard deviation of the ij'th protein fold-change and it is modeled to come from a distribution of errors centered at τ. This yields a partially pooled variance estimate for each protein fold-change, which provides an effective solution for estimating the variance of proteins with only a small number of observations.

The prior distribution $\beta_{ij} \sim N(0, 10)$ was selected to be weakly informative. The sampling is efficient since it does not have to consider impossible outcomes, but the prior will have virtually zero effect on the posterior estimate since a Gaussian distribution centered at zero with a standard deviation of 10 covers a range of values greatly exceeding the dynamic range of a TMT proteomics experiment (even infinite changes usually appear to have log2 fold-changes less than 6 or 7 (O'Brien et al. 2018)). Similarly, τ, which represents the average experimental error, has a half-normal prior distribution that far exceeeds the full range of plausible experimental error while preventing model instability caused by sampling unrealistic values. We use an inverse-gamma distribution with shape parameter of 2 in order to create a prior that enables stable sampling while covering all of the realistic possibile errors for an experiment with values that tend fall exclusively within (-10, 10). The shape parameter of 2 prevents sampling of errors near zero which are unrealistic and can result in unstable sampling, it also has the convienient property that the scale parameter, τ reprensents the mean of the distribution. Some of the properties of this prior distribution have been previously explored (Chung et al. 2013).

Partial pooling of the variance parameters results in estimates that converge to the within group (protein/condition combination) variance estimate as the number of observations increases. As the number of observations converges to zero the estimate is pulled towards the overall experimental error, τ. When there is only one observation we have noticed some instability in these estimates. For this reason we have programmed the model to directly use τ as the standard deviation for proteins with only a single peptide.

This Bayesian model is fit with the Stan programming languange (Carpenter et al. 2017). The reported estimate and variance for each protein are calculated as the mean and variance of the relevant posterior distribution.

Second Model

The above model provides estimates of log-fold changes for each protein from a reference to each condition in the experiment. Each of these estimates has variability from both within-sample peptide level variance and between sample variation across replicates. In many experimental designs the variation between biological replicates can large and separating out this source of heterogeneity can be highly informative. For this reason, by default, we fit two models. The above model which estimates log2 fold-changes between conditions and another which estimates log2 fold-changes between biological replicates. Estimates from the condition model are shown in the Precicion Plots, while Proportion Plots are generated by taking the ALR^{-1} of the posterior distributions from the model that estimates changes for each biological replicate.

It should be noted that when analyzing D dimensional compositions, the covariance matrices are degenerate because there are only $D-1$ degrees of freedom. Consequently, variance estimates for the proportions obtained on each of D biological replicates should be interpreted with caution. For each of the $D-1$ replicates that had a posterior distributions in log-ratio space, the variance estimates come from a 1-1 transformation of the posterior distribution and should be reliable. However, it is difficult to interpret the credible intervals on the new variables (usually proteins from the bridge channel). We considered dropping this interval from the plots, however we have found that variation in the bridge across plexes can result in larger than usual credible intervals. While these do not have the same interpretation as the other replicates, they can still be informative.

References

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data.* Chapman; Hall.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan : A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. doi:10.18637/jss.v076.i01.

Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika.* doi:10.1007/s11336-013-9328-2.

O'Brien, Jonathon J., Jeremy D. O'Connell, Joao A. Paulo, Sanjukta Thakurta, Christopher M. Rose, Michael P. Weekes, Edward L. Huttlin, and Steven P. Gygi. 2018. "Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags." *Journal of Proteome Research* 17 (1). American Chemical Society: 590–99. doi:10.1021/acs.jproteome.7b00699.